Towards Standards and Guidelines for Developing Open-Source and Benchmarking Learning for Robot Manipulation in the COMPARE Ecosystem

Adam Norton and Holly Yanco
University of Massachusetts Lowell
Lowell, MA, USA
adam_norton,holly_yanco@uml.edu

Ricardo Digiovanni Frumento and Yu Sun
University of South Florida
Tampa, FL, USA
ricardod,yusun@usf.edu

Kostas Bekris
Rutgers University
New Brunswick, NJ, USA
kostas.bekris@cs.rutgers.edu

Berk Calli
Worcester Polytechnic University
Worcester, MA, USA
bcalli@wpi.edu

Aaron Dollar
Yale University
New Haven, CT, USA
aaron.dollar@yale.edu

Abstract: The Collaborative Open-source Manipulation Performance Assessment for Robotics Enhancement (COMPARE) ecosystem will enable the robot manipulation community to more effectively conduct research and evaluate system performance, with the goal of enabling the quantitative comparison of approaches. COMPARE will addresses pertinent issues in robot manipulation (e.g., modularity of software, quality control, and testing frameworks), conducting outreach to build participation, and activating the ecosystem through activities such as workshops and competitions. Given the diversity of open-source products available for robot manipulation – including perception and planning packages, learning algorithms, datasets, benchmarking protocols, object sets, and hardware designs – the goal of the ecosystem is to create a greater cohesion and compatibility between these products via community-driven standards that will allow increased modularity and easier implementation of these products that enables enhanced performance quantification. This paper provides a brief overview of COMPARE and initial considerations for its efforts to improve robot learning.

Keywords: open-source, benchmarking, standards

1 Introduction

The inability to reliably grasp and manipulate objects in unstructured environments is a major limitation preventing fully assistive robot systems in the home for our growing elderly population, flexible manufacturing to improve the supply chain, and many other applications. Perhaps the biggest barrier to progress is the lack of an ability to quantitatively evaluate performance. While it is easy enough to specify details about standardized tasks that researchers might be asked to perform with their systems (e.g., how fast can you unpack a box of standard objects), the primary challenge is in doing so in a way that gives insight into exactly what worked — or did not — and why. Getting a robot system to perform such a manipulation task involves an enormous number of subsystems (e.g., perception, motion planning, learning), which most researchers generally only contribute to one at a

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

time. However, the other subsystems are generally not standardized, and each research group often crafts their own custom software stack (sometimes utilizing ROS [1], Move It [2], and other open-source products). As a result, it is nearly impossible to pinpoint the sources of success or failure for a specific system and to directly compare different systems.

After leading a series of workshops and surveys for the robot manipulation community [3], we have identified the needs of the community and developed a concept to create an open-source ecosystem (OSE) around many of the existing open-source products that are already available for robot manipulation (e.g., software components, datasets, benchmarking protocols, object sets, etc.). The goals of the COMPARE ecosystem are to (1) create a greater cohesion between open-source products by generating community-driven standards for components of software pipelines that increase modularity and enable implementation that allows for greater performance quantification, (2) unite the existing community of users and developers to build upon and integrate existing open-source products as well as improve and evolve the future of robot manipulation research, and (3) facilitate sufficient commonality in hardware and software that allows for quantitative evaluation of research and eases the implementation of the complex robot manipulation pipeline.

This paper briefly introduces the COMPARE ecosystem and initial considerations for improving robot learning for generalization and benchmarking.

2 Scope of the Open-Source Ecosystem

This new community-driven OSE – COMPARE – will support robot manipulation research and benchmarking, seeking to improve the compatibility, modularity, interoperability, and replicability of open-source products used in robot manipulation including software components (e.g., algorithms and packages), benchmarking protocols, objects and artifacts, datasets, and hardware designs, for more effective, informative, and reproducible benchmarking. The goal is ultimately to move more rapidly towards finding solutions to the open problem of robotic perception, grasping, and manipulation of a wide variety of objects, tasks, and applications. It will lead to the development of a community of contributors and users, expanding collaborations and the number of people who can participate in this challenging research area. We will address issues related to standards and guidelines for the ecosystem to adopt with regards to the open-source products it supports, as well as the ways those products integrate into larger frameworks for implementing and controlling complete robot manipulation pipelines. Robot learning is one of many components within the open-source and benchmarking for robot manipulation ecosystem that COMPARE aims to improve.

3 Community Needs

We engaged with the open-source robotics community through several mechanisms: an online survey (with over 100 respondents), four workshops at robotics conferences [3], and over 60 deep dive meetings with key community stakeholders. Through these activities, our team gathered substantial input regarding the current landscape in terms of limitations and opportunities, how the OSE should be structured, activities it should conduct, and best practices it should follow, as well as began informally building the community and exposing the idea of developing the OSE.

Current limitations include a lack of relevant comparable benchmarks, limited simulation capabilities, and issues when integrating open-source products, and respondents indicated that they did not frequently contribute to open-source or benchmark to compare to others in the field. The lack of clear instructions in published benchmarks was also highlighted. Several bottlenecks to conducting quality robot manipulation performance evaluation were identified, including variation in available resources across labs for physical robot testing (robots, personnel, tools), lack of community consensus on several topics (metrics, protocols, data collection methods, etc.), lack of truly modular software to enable component-level and holistic system evaluations, and the lack of incentives as publications do not require/favor benchmarking.

Recommendations for improvement include organized repositories of robot manipulation benchmarking results and open-source products for robot manipulation, and the benefit of having truly modular software to enable improved performance benchmarking was also highlighted. Recommendations of OSE activities in order to build incentives for the community to conduct quality robot manipulation performance evaluation include advocating for publication review criteria to favor research that includes benchmarking comparisons and utilization of open-source products, developing mechanisms to ensure industry relevance and applicability of benchmarks, and set performance targets for benchmarking across the community (e.g., a competition) by establishing a desired threshold of performance rather than just relative performance comparison.

4 Structure and Activities

COMPARE utilizes an organization and governance structure that consists of a **steering committee** for roadmap development and prioritization, an **advisory board** for oversight and sustainability counsel, a **community facilitator** for community interaction and representation, and **working groups** comprised of developers and users to establish open-source product development guidelines and standards. The initial set of broad working groups (WG): Software Components WG, Benchmarking Protocols WG, Objects and Artifacts WG, Datasets WG, and Hardware Designs WG. Within each WG, narrower task groups (TG) are in process of being defined, such as TGs for Perception, Grasp Planning, and Learning under the Software Components WG.

Using this structure, the OSE will produce **standards and guidelines** for open-source product development, conducting benchmarking, and reporting performance results. **Online resources** are under development that will include communication channels such as discussion forums, repositories of open-source products for robot manipulation, and leaderboards of benchmarking results (similar to Papers With Code [4]). **Test and evaluation procedures** will be developed to validate the open-source products and their compatibility with others in the software pipeline, including reviews of their adherence to the developed standards and guidelines, performance assessments of their functionality, and usability analyses when attempting to integrate the products.

A set of events are planned in order to jumpstart activity in the OSE and engage with the community users and developers. These events will focus the community around a series of salient topics in order to share their research, develop best practices, and integrate their open-source products together. Activities include **conference workshops** on benchmarking protocols and software pipelines, **seasonal schools** to develop and integrate the open-source products, and **competitions**.

See Figure 1 for a diagram of the open-source development and performance evaluation processes correlated to COMPARE infrastructure to support it.

5 Towards Improving Robot Learning in COMPARE

This section very briefly outlines the initial high-level directions of the COMPARE efforts that will address robot learning. It is by no means a comprehensive characterization of the problem space, an effort which is currently underway in order to better inform the initial WG/TG structure.

The topic of robot learning will be discussed in several of the COMPARE WGs/TGs, including **standards and guidelines for best practices in developing open-source software components that can be more easily integrated and adapted by users, both in terms of training learning algorithms and executing tasks**. For instance, during our community outreach and research, example implementations of algorithms was highlighted as a strongly desired feature of open-source software components. Additionally, new peer review processes for quality control of contributed open-source were also suggested. So, to meet the to-be-developed COMPARE standards, such features of contributed open-source for learning may be required. These are only examples used to illustrate how the goals and outputs of COMPARE are driven by community input; actual development of the standards and guidelines is yet to be underway.

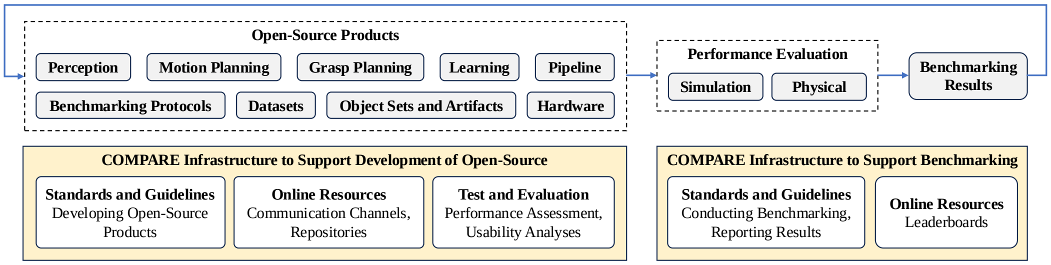

Figure 1: Diagram of open-source product development and conducting performance evaluations (top) and the supporting COMPARE infrastructure (bottom).

Benchmarking of robot learning is also of significant importance to COMPARE. The goal is to establish clear standards and guidelines that drive advancements in robot learning, particularly in the areas of generalization and transfer learning across different robotic platforms and environments. One of the main challenges in this domain is ensuring that robots can transfer learned skills to new, unseen tasks and physical setups as described in [5]. To accomplish such an endeavor, it is necessary to use an established protocol to evaluate a robot's ability to generalize across various configurations, hardware, and tasks, which requires robust benchmarking that addresses both learning methods and the physical aspects of robots, such as sensor calibration, actuation differences, and task variations. **COMPARE standards and guidelines may set minimum requirements for these aspects to qualify as an acceptable robot learning benchmark**.

The initial COMPARE efforts will build on existing benchmarks like [6], [7], [8] and community engagement to propose novel evaluation methodologies that emphasize task variability, environmental dynamics, and robot morphology. **We recommend introducing more systematic variation in task complexity, sensor inputs, and physical conditions**, as we aim to test a robot's learning adaptability and its ability to perform under diverse real-world constraints. This includes testing for sim-to-real transfer and adapting to hardware differences, which remain key bottlenecks in robotic manipulation. Feedback on additional salient factors for enhanced benchmarking is sought.

Incorporating internet-scale data, such as unstructured videos [9] from platforms like YouTube, into our benchmarks opens new avenues for large-scale, self-supervised learning. Robots can leverage this data to learn general-purpose skills by observing a wide range of human activities, tasks, and environmental interactions. However, using this data requires developing techniques for filtering, annotating, and structuring unstructured video data to make it usable for robot learning models. The COMPARE WGs/TGs aim to **explore the challenges of leveraging noisy, uncontrolled data and propose ways to integrate this resource into structured benchmarks that support multi-task learning**.

Finally, COMPARE efforts may also address the current limitations of the field by offering a comprehensive task set that spans multiple domains (e.g., manipulation of rigid and deformable objects, human-robot interaction) which was attempted in [10]. We will propose standardized protocols and performance metrics that allow researchers to test generalization across unseen tasks and new robots, ensuring that the benchmarks reflect real-world conditions. By doing so, **we aim to provide quantifiable metrics that allow for reliable cross-platform comparisons, which are currently missing in many robot learning benchmarks**.

6 How To Participate

We are currently soliciting input as to how such an OSE for improving robotic manipulation research and benchmarking should be developed, organized, and run. Additionally, we welcome participation from any and all robotic manipulation researchers to contribute to the OSE activities, participate in working groups, and attend the planned events. For more information, contact the lead author to express your interest in participating and visit the COMPARE website [11].

Acknowledgments

This work was supported in part by the National Science Foundation (awards TI-2346069 and TI-2229577).

References

[1] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*, volume 3, page 5. Kobe, Japan, 2009.

[2] MoveIt Motion Planning Framework. https://moveit.ros.org/. Accessed: 2023-08-04.

[3] COMPARE - Background. https://www.robot-manipulation.org/background. Accessed: 2024-09-24.

[4] The Latest in Machine Learning - Papers With Code. https://paperswithcode.com/. Accessed: 2024-09-24.

[5] B. Calli, A. Dollar, M. A. Roa, S. Srinivasa, and Y. Sun. Guest editorial: Introduction to the special issue on benchmarking protocols for robotic manipulation. *IEEE Robotics and Automation Letters*, 6(4):8678–8680, 2021.

[6] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, et al. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9197–9203. IEEE, 2023.

[7] O. Ahmed, F. Träuble, A. Goyal, A. Neitz, Y. Bengio, B. Schölkopf, M. Wüthrich, and S. Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

[8] X. Liu, F. Wan, S. Ge, H. Wang, H. Sun, and C. Song. Jigsaw-based benchmarking for learning robotic manipulation. In *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 124–130. IEEE, 2023.

[9] C. Eze and C. Crick. Learning by watching: A review of video-based learning approaches for robot manipulation. *arXiv preprint arXiv:2402.07127*, 2024.

[10] C. Chamzas, C. Quintero-Pena, Z. Kingston, A. Orthey, D. Rakita, M. Gleicher, M. Toussaint, and L. E. Kavraki. Motionbenchmaker: A tool to generate and benchmark motion planning datasets. *IEEE Robotics and Automation Letters*, 7(2):882–889, 2021.

[11] Collaborative Open-source Manipulation and Perception Assets for Robotics Ecosystem (COMPARE). https://www.robot-manipulation.org/. Accessed: 2024-09-24.